3В

УДК 004.934

Н.Б. Васильєва
Міжнародний науково-навчальний центр інформаційних технологій та систем, м. Київ, Україна
Україна, 03680, просп. Акад. Глушкова, 40, МСП, м. Київ, n.vassilleva@gmail.com, ninel@uasoiro.org.ua

Дослідження невідповідності шкали акустичної та лінгвістичної моделей розпізнавання злитого українського мовлення

N.B. Vasylieva
Department of Speech Recognition and Synthesis, International Research and Training Center of Information Technologies and Systems, c. Kiev, Ukraine
Ukraine, 03680, Acad. Glushkov Ave., 40, MSP, Kyiv, n.vassilleva@gmail.com, ninel@uasoiro.org.ua

Exploration of Acoustic and Linguistic Models Scale Discrepancy for Continuous Ukrainian Speech Recognition

Н.Б. Васильева
Международный научно-учебный центр информационных технологий и систем, г. Киев, Украина
Украина, 03680, ул. Акад. Глушкова, 40, МСП, г. Киев,n.vassilleva@gmail.com, ninel@uasoiro.org.ua

Исследование несоответствия шкалы акустической и лингвистической моделей распознавания слитой украинской речи

У статті описується розробка експериментальної системи перетворення мовленнєвого сигналу на текст, що складається як зі слів, так і з субслівних елементів. Велику увагу приділено вибору навчальної вибірки для оцінки параметрів акустичної моделі розпізнавання. Зокрема розглядалися такі варіанти: модель, побудована лише на злитому мовленні; модель, що об’єднує злите мовлення та ізольовані слова; модель, що не враховує наголошеність голосних; та модель, що враховує наголошеність лише голосних «и» та «е». Проводиться оцінка параметрів акустичної моделі на основі однодикторного мовленнєвого корпусу. Вибираються коефіцієнти, які компенсують невідповідності шкали акустичної та лінгвістичної складової моделі розпізнавання. Наводяться результати експериментальних досліджень.
Ключові слова: під-слово, склад, розпізнавання мовлення, навчальна і контрольна вибірки, злите мовлення, ізольовані слова.

This paper describes the development of experimental systems of speech signal to text conversion based on words and sub-words. Main attention is paid to selecting of training set for estimation of the parameters of acoustic recognition models. Particularly, the following options are considered: acoustic model based only on continuous speech, a model that integrates continuous speech and isolated words, a model that ignores stress vowels, and a model that takes into account only stress vowels “y” and “e”. The estimation of acoustic model parameters is based on mono-speaker speech corpus. The factors compensating the inconsistency of acoustic and linguistic component model scales are analyzed and their values are explored. The results of experimental research are discussed.
Key Words: sub-word, syllable, speech recognition, training and test sets, continuous speech, isolated words.

В статье описывается разработка экспериментальной системы преобразования речевого сигнала в текст, который состоит как из слов, так и из субсловных элементов. Большое внимание уделено выбору обучающей выборки для оценки параметров акустической модели распознавания. В частности рассматривались такие варианты: акустическая модель, построенная только на слитной речи; модель, объединяющая слитую речь и изолированные слова; модель, не учитывающая ударность гласных; и модель, учитывающая ударность только гласных «ы» и «е». Проводится оценка параметров акустической модели на основе однодикторного речевого корпуса. Выбираются коэффициенты, компенсирующие несоответствия шкалы акустической и лингвистической составляющей модели распознавания. Приводятся результаты экспериментальных исследований.
Ключевые слова: под-слова, слог, распознавание речи, обучающая и контрольная выборки, слитная речь, изолированные слова.

Вступ

Системи пофонемного розпізнавання зазвичай оперують алфавітом фонем (контекстно залежних або контекстно незалежних), з яких складаються мовленнєві образи слів. Потім на слова накладаються обмеження їх слідування шляхом побудови лінгвістичної моделі (ЛМ) або граматик. При збагаченні лексики зростають обсяги робочого словника, суттєво ускладнюються граматика або ЛМ, а це призводить до зменшення продуктивності системи розпізнавання.

Якщо використовувати замість слів мовленнєві образи складів або морфем, то збагачення лексики не призведе до помітного зростання робочих словників та ускладнення граматики чи ЛМ. При цьому постає проблема переходу від послідовностей складів (морфем) до послідовностей слів, оскільки помилка розпізнавання складу або морфеми може спричиняти ситуацію, коли їх послідовностям не можливо безпосередньо зіставити слово.

У попередній роботі [1] досліджувалась надійність розпізнавання фонем і складів двох видів. Для проведення експериментальних досліджень використовувався однодикторний мовленнєвий корпус злитого мовлення. Велику увагу приділено створенню навчальної вибірки (НВ): вибору початкового текстового корпусу (ТК), алгоритму вибору текстів, оброблення «Жадібним» алгоритмом вибраних текстів, запису мовленнєвої НВ. Алфавіт корпусу НВ налічував близько 51 тис. фонем-трифонів у 18 тис. реченнях. Обсяг словника – 47,5 тис. слів. Загальна кількість реалізацій слів у цій НВ – 184,9 тис. Отримано близько 36 годин запису акустичної бази навчальної вибірки. Також був описаний алгоритм вибору НВ для ізольованих слів. Розглядалися два словника: словник УМІФ та словник частотних слів. Обсяг словника НВ ізольованих слів склав ~ 13 тис. слів та після запису більш як 12 годин мовлення.

Графіки частоти фонем-трифонів у різних джерелах (текстовий корпус, словник УМІФ та частотний словник) та отримані відповідні НВ наведені на рис 1. Тут можна побачити, що при роботі «Жадібного» алгоритму кількість елементів, що зустрілися один раз, збільшилися в декілька разів для кожної НВ. Також з рисунка випливає, що частота фонем-трифонів загалом відповідає розподілу Ципфа – Мандельброта як для вхідних корпусів, так і після роботи «Жадібного» алгоритму.

Експерименти проводилися на різних контрольних вибірках (КВ). Перша КВ формувалася за принципом частотності фонем-трифонів, що використовуються. «Частотна» КВ складалася з 3 тис. речень, обсяг словника мав 3 225 слів, загальна кількість реалізацій яких – 8 987 слів. Отримана КВ має 3,6 годин запису. Друга КВ вибиралася випадковим чином з тих самих текстів, з яких вибирався текст НВ. «Випадкова» КВ складалася з 2 тис. речень, обсяг словника мав 10 013 слів, загальна кількість реалізацій яких – 22 864 слів. Отримана КВ має 4,3 годин запису. Третя КВ була вибрана з

текстів, які не використовувалися для вибору НВ. Для цього із сайта української Вікіпедії [2] випадковим чином вибраний зв'язний текст (1,2 тис. речень). Обсяг словника складав 7,3 тис. слів. Загальна кількість реалізацій слів – 16 тис.

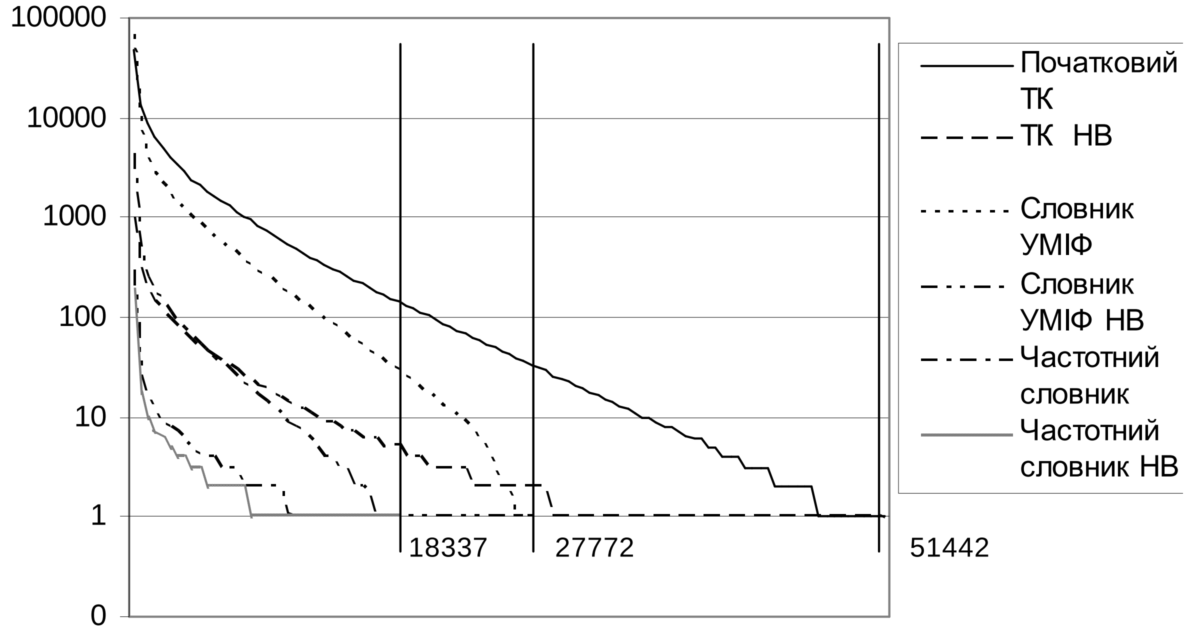


Рисунок 1 – Розподілення фонем-трифонів за частотністю в текстових вибірках

Процедура розпізнавання проводилась за допомогою декодерів *HTK* [3] і *Julius* [4] на трьох КВ: частотній, випадковій та «Вікіпедія». Як елемент робочого словника бралися: фонеми (всього 59), відкриті склади (всього 7 270), склади, поділені за правилами складоподілу (всього 10 200) та цілі слова.

Метою даної роботи є дослідження невідповідності акустичної та лінгвістичної компонент математичної моделі розпізнавання мовленнєвого сигналу з метою підвищення надійності розпізнавання.

В наступному розділі описується обґрунтування створення різних акустичних моделей. Потім приділяється увага вибору коефіцієнтів, які компенсують невідповідності шкали акустичної та лінгвістичної складової моделі розпізнавання. Проводяться експериментальні дослідження, наводяться та обговорюються результати досліджень.

Побудова експериментальної системи перетворення мовлення на текст

Проводилося оцінювання параметрів акустичних моделей з використанням програмного інструментарію *HTK* та *Julius*. Акустичні моделі формувалися на основі контекстно незалежних фонем, оскільки їх алфавіт невеликий, а отже, для статистичних оцінок необхідна менша база акустичних сигналів, ніж для складів і фонем-трифонів, яких більше в тисячі разів, і топологія їх акустичних моделей вимагає додаткових досліджень. Порівняно з попередньою роботою, в якій акустична модель будувалася лише на базі злитого мовлення, в даній розглядалася також модель фонем, побудована як на злитому мовленні, так і на ізольованих словах. Це зроблено з метою покращення розпізнавання для «Частотної» КВ, яка давала найгірші результати з обраних КВ.

У табл. 1 представлені результати фонемної помилки розпізнавання при використанні різних акустичних моделей. За результатами, наведеними в табл. 1, видно, що застосування акустичної база НВ ізольованих слів (ІС) на додачу до НВ злитого мов-

лення (ЗМ) приводить до зменшення фонемної помилки. Це можна пояснити тим, що акустична база ІС збільшує кількість реалізацій кожної фонеми. Також наявність коротких синтагм (що характерно для природного людського мовлення) сприяє покращенню результатів розпізнавання.

Таблиця 1 – Показники фонемної помилки розпізнавання – *PER* (%) для КВ злитого мовлення на основі різних мовленнєвих образів, з використанням різних акустичних моделей (злитого мовлення та ізольованих слів) інструментарієм *HTK*

Назва контрольної вибірки	Фонема		Відкритий склад		Склад, ділений за правилами складоподілу	
	акустична модель, що використовувалась					
	ЗМ	ЗМ + ІС	ЗМ	ЗМ + ІС	ЗМ	ЗМ + ІС
«Випадкова» КВ	28,86	25,6	24,92	23,06	24,54	**21,34**
«Випадкова» КВ (без наголосу)	21,39	20,96	17,68	**17,00**	17,29	18,36
«Частотна» КВ	36,6	26,7	37,75	23,22	-	**22,33**
«Частотна» КВ (без наголосу)	26,1	21,56	27,95	**17,49**	-	18,11
КВ «Вікіпедія»	31,93	30,76	**28,01**	28,53	28,18	29,35
КВ «Вікіпедія» (без наголосу)	24,72	24,52	28,81	22,02	**21,00**	22,88

Для кожної з 57 фонем української мови і двох фонем-пауз отримані моделі, які мають кожна три стани та від 4 до 36 сумішей нормальних законів залежно від частотності.

Акустичні моделі для розпізнавання будувалися, враховуючи наголошені голосні. На письмі ж наголос зазвичай опускається. Виходячи з цих міркувань наголос не враховувався, що дало значно меншу оцінку *PER*. А чи вплине на надійність розпізнавання, якщо в акустичній моделі не враховувати наголошеність голосних? Щоб дослідити це, була створена акустична модель, яка прирівнює наголошені та ненаголошені відповідні голосні: і+ до і, а+ до а, о+ до о, и+ до и, е+ до е, у+ до у. Також, щоб врахувати специфіку українського мовлення, а саме редукцію ненаголошених е, и до $и^е$, $е^и$ та навпаки, була створена акустична модель, в якій залишилися лише дві наголошені голосні е+ та и+.

Компенсування невідповідності шкали акустичної та лінгвістичної складових моделі розпізнавання

Декодер намагається знайти послідовність слів або їх компонент $\mathbf{q}_{1:L} = \mathbf{q}_1,...,\mathbf{q}_L$, які найбільш правдоподібно генерують послідовність векторів, що спостерігаються $\mathbf{Y}_{1:T} = \mathbf{y}_1,...,\mathbf{y}_L$, виходячи з інтегральної міри схожості:

$$\hat{\boldsymbol{q}} = \arg\max_{\mathbf{q}}\{\log p(\mathbf{Y}\,|\,\mathbf{q}) + (\alpha\,\log(P(\boldsymbol{q})) + \beta\,|\,\mathbf{q}\,|)\},$$

де α та β – коефіцієнти, які компенсують невідповідності шкали акустичної моделі (АМ) та лінгвістичної моделі (ЛМ), які є компонентами математичної моделі автоматичного розпізнавання мовленнєвого сигналу. Тому на першому етапі проводилися експерименти з ціллю емпірично підібрати параметри α та β, рекомендований діапазон яких складає 0 – 20 та 0 – (–20) відповідно [3], [5].

При оцінці надійності використовувались показники фонемної помилки (англійською, *PER – Phoneme Error Rate*):

$$\%PER = 100\% - \frac{H - I}{N} 100\%$$

та фонемної некоректності (*PIR – Phoneme Incorrectness Rate*):

$$\%PIR = 100\% - \frac{H}{N} 100\%,$$

де H – кількість правильно розпізнаних під-слівних елементів;

I – кількість помилково вставлених під-слівних елементів;

N – загальна кількість промовлених під-слівних елементів.

Проводилися експериментальні дослідження корегування шкали невідповідності АМ та ЛМ для слів і їх компонентів.

Дослідження невідповідності шкали АМ і ЛМ для під-слівних елементів. На рис. 2 – 7 проілюстровані показники % *PER* та % *PIR* фонемного розпізнавання згаданих вище КВ при змінах коефіцієнта β в п’яти точках (0, –5, –10, –15, –20) для α, що дорівнює 0, 5 та 10.

Зменшення *PER* відбувається головним чином за рахунок скорочення кількості вставлених під-слівних елементів, яких не має бути. Ріст некоректності обумовлений зменшенням правильно розпізнаних елементів. З рис. 2 – 7 слідує, що найменша фонемна помилка досягається при значеннях параметрів α = 5 та β = –5. Показник коректності *PIR* дав можливість визначити, що надійність виросла за рахунок скорочення числа вставок.

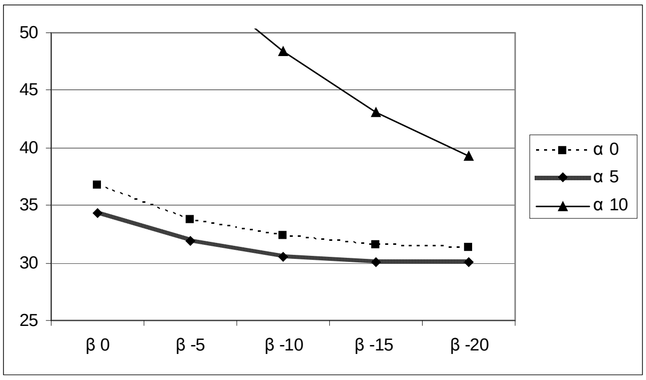


Рисунок 2 – Показники *PER* розпізнавання (%) для злитого мовлення на «Частотній» КВ

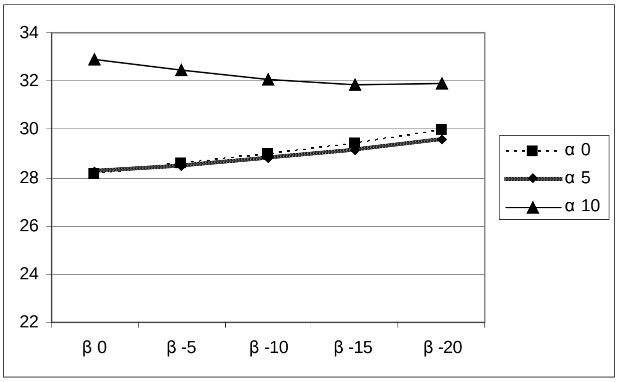


Рисунок 3 – Показники *PIR* розпізнавання (%) для злитого мовлення на «Частотній» КВ

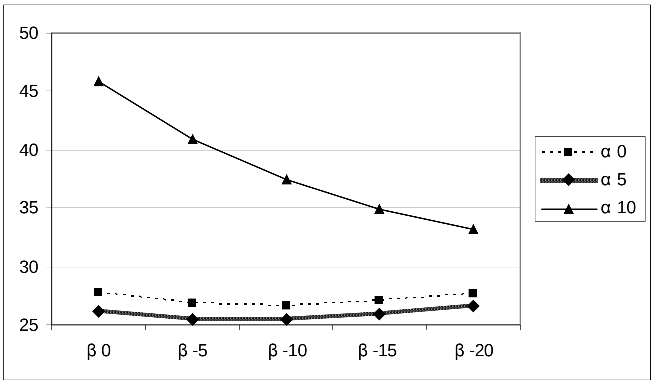


Рисунок 4 – Показники *PER* розпізнавання (%) для злитого мовлення на «Випадковій» КВ

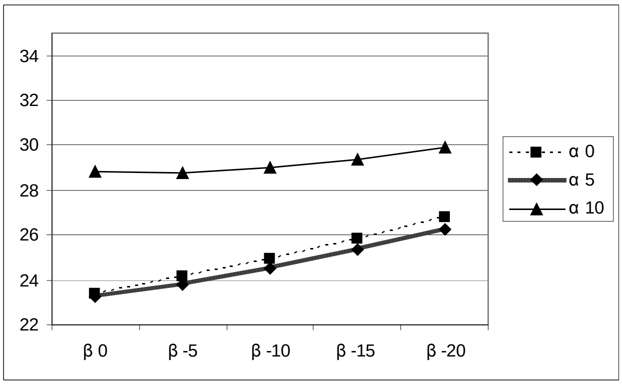


Рисунок 5 – Показники *PIR* розпізнавання (%) для злитого мовлення на «Випадковій» КВ

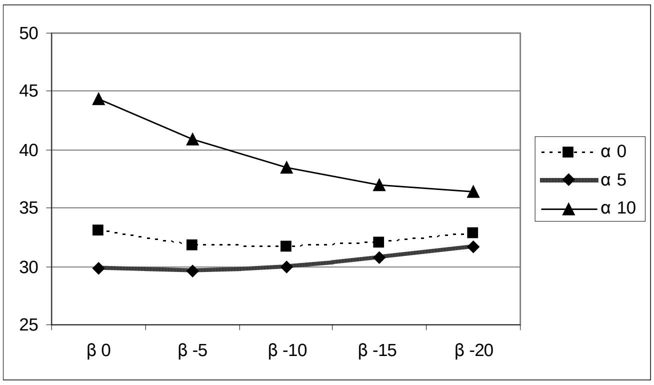


Рисунок 6 – Показники *PER* розпізнавання (%) для злитого мовлення на КВ «Вікіпедія»

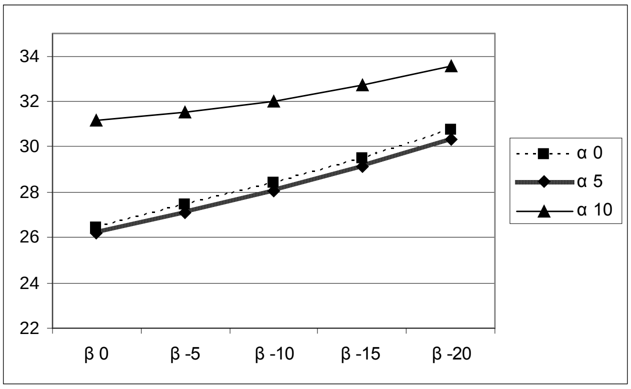


Рисунок 7 – Показники *PIR* розпізнавання (%) для злитого мовлення на КВ «Вікіпедія»

Дослідження невідповідності шкали АМ і ЛМ для слів. Проводилися експериментальні дослідження послівного розпізнавання злитого мовлення із застосуванням ЛМ. На рис. 8 – 10 наведені результати послівного розпізнавання злитого мовлення для різних КВ при змінах коефіцієнта β в чотирьох точках (0, –1, –2, –5) для α, що дорівнює 4, 7 та 15. Словник лінгвістичної моделі розпізнавання складав 100 тис. слів. При цьому слів, яких немає в словнику (*Out of Vocabulary – OOV*), було 2,3% для «Частотної» КВ, 9,6% для «Випадкової» КВ та 7% для КВ «Вікіпедія». Якщо врахувати попередні дослідження (табл. 1), то очевидно для навчання АМ потрібно брати обидві навчальні вибірки акустичних баз: злитого мовлення та ізольованих слів. На рис. 8 – 10 вони позначені як *TC_SL*; позначення *TC_SL_(EY)* має акустична модель, побудована на НВ злитого мовлення та ізольованих слів, які розрізняють наголошеність голосних лише для *и* та *е*; позначення *TC_SL_beznag* має акустична модель, побудована так само, як і попередні, але не розрізняє наголошеність голосних.

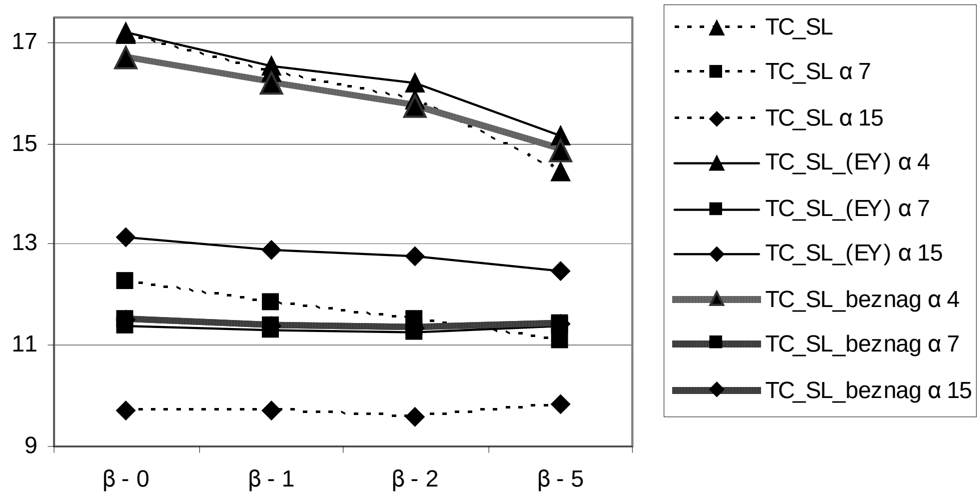


Рисунок 8 – Показники помилки надійності послівного розпізнавання (%) для злитого мовлення на «Частотній» КВ

Показник послівної некоректності так само, як і для фонемного розпізнавання, зростає при збільшенні штрафів за рахунок випадання елементів, які мають бути. Як видно з рис. 8 – 10, коефіцієнт α = 15 та β = -2 з акустичною моделлю, яка враховує всі наголошені голосні, поводить себе краще, ніж інші, для всіх КВ. Графіки *TC_SL_beznag* α = 7 та *TC_SL_beznag* α = 15 (рис. 8 – 10) мають майже однакові траєкторії, які на рисунках зливаються в одну.

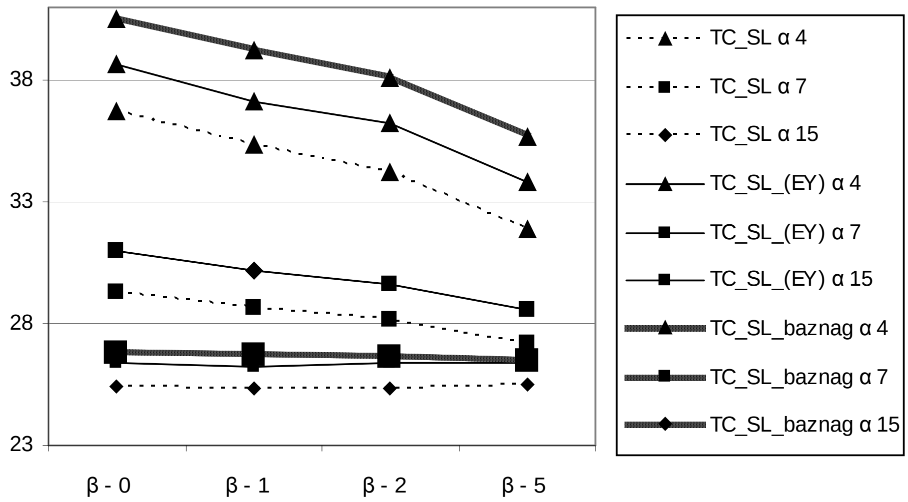


Рисунок 9 – Показники помилки надійності послівного розпізнавання (%) для злитого мовлення на «Випадковій» КВ

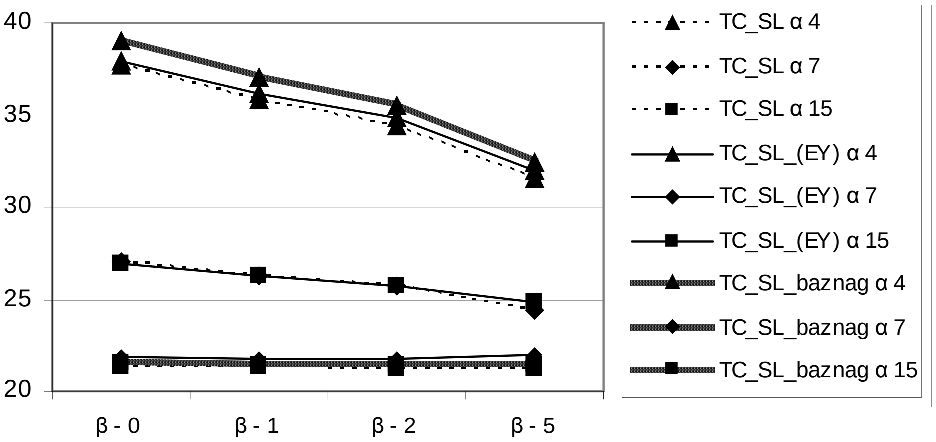


Рисунок 10 – Показники помилки надійності послівного розпізнавання (%) для злитого мовлення на КВ «Вікіпедія»

Висновки

Як і планувалося в попередній роботі [1], для покращення розпізнавання злитого мовлення були задіяні обидві акустичні НВ: злитого мовлення та ізольованих слів. Це значно покращило результати надійності розпізнавання, особливо на «Частотній» КВ.

У наведених експериментах допускалася вільна граматика слідування під-слівних елементів, а для слів використано статистичні лінгвістичну модель.

Велику увагу в даній роботі було приділено дослідженню та підбору коефіцієнтів α та β, які компенсують невідповідності шкали акустичної та лінгвістичної складової моделі розпізнавання. З однієї сторони, збільшення значення коефіцієнтів впливає на кількість розпізнаних під-слівних елементів, зменшуючи коректність розпізнавання, а з іншої сторони – на правильність розпізнаних елементів, збільшуючи надійність.

Планується застосувати статистичні лінгвістичні моделі для під-слівних елементів, що має привести до зменшення помилки розпізнавання. Залишається недослідженим вплив багатьох параметрів декодерів на надійність та швидкість. Зокрема, будуть розроблятися підходи до зменшення алфавіту складів, що має прискорити розпізнавання.

Література

1. Васильєва Н.Б. Використання граматик вільного порядку слідування фонем і складів для пофонемного розпізнавання злитого мовлення / Н.Б. Васильєва // Штучний інтелект. – 2011. – № 4. – С. 80-86.
2. Режим доступу : http://uk.wikipedia.org

3. HTK Book, version 3.1 / Young S.J. et al. – Cambridge University, 2002.
4. Lee A. Julius – an open source real-time large vocabulary recognition engine / A. Lee, T. Kawahara and K. Shikano // In Proc. European Conference on Speech Communication and Technology (EUROSPEECH), 2001. – P. 1691-1694.
5. Gales Mark. The Application of Hidden Markov Models in Speech Recognition / Mark Gales, Steve Young // Foundations and Trends in Signal Processing. – 2007. – Vol. 1, № 3. – P. 195-304.

Literatura

1. Vasylieva N. Shtuchnyy Intelekt. Donec'k. 2011. № 4. S. 80-86.
2. http://uk.wikipedia.org
3. Young S.J. et al., HTK Book, version 3.1, Cambridge University, 2002.
4. A. Lee, T. Kawahara and K. Shikano: Julius – an open source real-time large vocabulary recognition engine. Eurospeech'2001, pp. 1691-1694.
5. Mark Gales, Steve Young. The Application of Hidden Markov Models in Speech Recognition. Foundations and Trends in Signal Processing Vol. 1. No. 3 (2007). P. 195-304.

RESUME

N.B. Vasylieva

Exploration of Acoustic and Linguistic Models Scale Discrepancy for Continuous Ukrainian Speech Recognition

This paper describes the development of experimental systems of speech signal to text conversion based on words and sub-words. The modification of the generative model known as Hidden Markov Model is applied to acoustic essence of phoneme.

Main attention is paid to selecting of training set for estimation of the parameters of acoustic recognition models. Particularly the following options are considered: an acoustic model based only on continuous speech, a model that integrates continuous speech and isolated words, a model in which the accented vowels are approximated with the corresponding unstressed vowels. That is *i*+, *a*+, *o*+, *e*+, *u*+, *y*+ are replaced respectively with *i*, *a*, *o*, *e*, *u*, *y*. Another model considers only two accented vowels *y*+ and *e*+. These allows taking into account some specifics of the Ukrainian language, namely the reduction of unstressed e and y to e^y, y^e and vice versa.

The estimation of acoustic model parameters is based on mono-speaker speech corpus. Testing was conducted on three control sets: "Frequency", "Random" and "Wikipedia". The factors compensating the inconsistency of acoustic and linguistic component model scales are analyzed and their values are explored. In these experiments a free order for sub-word elements is investigated and a statistical language model for words-by-word recognition is estimated. The experimental results show that correctly tuned scaling coefficients may significantly improve the recognition accuracy.

Стаття надійшла до редакції 02.07.2012.